Describing data set May 2020

Joost Bloos

06/11/2021

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

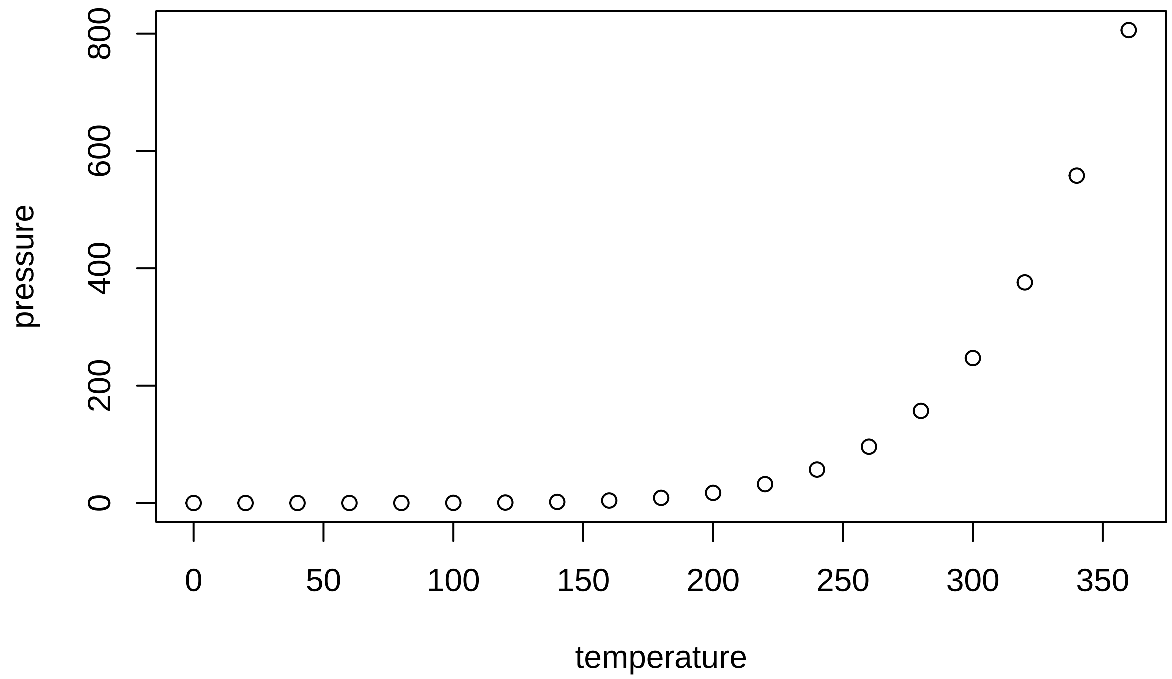
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
#install.packages("ISwR")
#install.packages("dplyr") # includes ggplot
#install.packages("ggplot")
#install.packages("ggplot2")
#install.packages("twitteR")
#install.packages("tidyr")
#install.packages("tidyverse")
#install.packages("ggmap")
#install.packages("sf")
#install.packages("mapview")
#install.packages("maps")
#install.packages("magrittr")
#install.packages("rgeos")
#install.packages("revgeo")
#install.packages("NLP")

#install.packages(c("cowplot", "googleway", "ggplot2", "ggplot", "ggrepel", "ggspatial", "libwgeom", "s


library(dplyr)
```

```
##
## Attaching package: 'dplyr'
## The following objects are masked from ’package:stats’:
##
##     filter, lag

## The following objects are masked from ’package:base’:
##
##     intersect, setdiff, setequal, union
```

```
library(tm)
```

```
## Loading required package: NLP
```

```
library(ISwR)
library(twitteR)
```

```
##
## Attaching package: ’twitteR’

## The following objects are masked from ’package:dplyr’:
##
##     id, location
```

```
library(tidyr)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x ggplot2::annotate() masks NLP::annotate()
## x dplyr::filter()     masks stats::filter()
## x twitteR::id()       masks dplyr::id()
## x dplyr::lag()        masks stats::lag()
## x twitteR::location() masks dplyr::location()
```

```
library(ggmap)
```

```
## Google’s Terms of Service: https://cloud.google.com/maps-platform/terms/.

## Please cite ggmap if you use it! See citation("ggmap") for details.
```

```
library(mapview)
library("rnaturalearth")
library("rnaturalearthdata")
library(devtools)
```

```
## Loading required package: usethis
```

```r
library(devtools)
install_github('mhudecheck/revgeo')
```

```
## Skipping install of 'revgeo' from a github remote, the SHA1 (5f01ff67) has not changed since last ins
##   Use `force = TRUE` to force installation
```

```r
library(revgeo)
```

```r
data1 <- read.csv("corona_tweets_59 May 2020", sep=",", stringsAsFactors = F, na.strings = c("","NA"))
```

```
## Warning in scan(file = file, what = what, sep = sep, quote = quote, dec = dec, :
## embedded nul(s) found in input
```

```r
#brief description of original hydrated data set May 2020:
str(data1) #shows number of observation out of 35 variables
```

```
## 'data.frame':    2274036 obs. of  35 variables:
##  $ coordinates               : chr  NA NA NA NA ...
##  $ created_at                : chr  "Sat May 16 04:13:37 +0000 2020" "Sat May 16 04:13:35 +0000 2020
##  $ hashtags                  : chr  "TheTourWillGoOn" NA NA "BREAKING COVID19" ...
##  $ media                     : chr  NA NA NA NA ...
##  $ urls                      : chr  "https://twitter.com/kennychesney/status/1261463380809809920" NA
##  $ favorite_count            : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ id                        : num  1.26e+18 1.26e+18 1.26e+18 1.26e+18 1.26e+18 ...
##  $ in_reply_to_screen_name   : chr  NA NA NA NA ...
##  $ in_reply_to_status_id     : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ in_reply_to_user_id       : num  NA NA NA NA NA NA NA NA NA NA ...
##  $ lang                      : chr  "en" "en" "en" "en" ...
##  $ place                     : chr  NA NA NA NA ...
##  $ possibly_sensitive        : chr  "false" NA "false" NA ...
##  $ quote_id                  : num  1.26e+18 NA NA NA NA ...
##  $ retweet_count             : int  0 7171 10 76 9 0 621 2 0 807 ...
##  $ retweet_id                : num  NA 1.26e+18 1.26e+18 1.26e+18 1.26e+18 ...
##  $ retweet_screen_name       : chr  NA "MSharifpourMD" "YahooFinance" "davejourno" ...
##  $ source                    : chr  "<a href=\"http://twitter.com/#!/download/ipad\" rel=\"nofollow\
##  $ text                      : chr  "I am sad, but I think itâ\200\231s the right decision. #TheTour
##  $ tweet_url                 : chr  "https://twitter.com/kevinhansford/status/1261510087874678784" "
##  $ user_created_at           : chr  "Thu May 14 15:39:20 +0000 2009" "Wed Oct 04 21:22:03 +0000 2017
##  $ user_id                   : num  4.00e+07 9.16e+17 3.69e+07 2.02e+07 2.42e+09 ...
##  $ user_default_profile_image: chr  "false" "false" "false" "false" ...
##  $ user_description          : chr  "Huge sports fan. Alabama Football, Kentucky Basketball and Gree
##  $ user_favourites_count     : int  8456 25418 103 156943 23317 12011 4243 67188 5349 9400 ...
##  $ user_followers_count      : int  52 87 1333 586 166 83 386 3975 2108 3108 ...
##  $ user_friends_count        : int  446 95 5002 499 763 206 2952 4648 3180 3678 ...
##  $ user_listed_count         : int  5 1 158 47 0 2 2 1 0 2 ...
##  $ user_location             : chr  NA "Formation" NA NA ...
##  $ user_name                 : chr  "Kevin â\200œThe Senatorâ\200\235" "BLM | The little garage" "pa
##  $ user_screen_name          : chr  "kevinhansford" "pbm_ssb" "chicago2503" "bigred_13" ...
##  $ user_statuses_count       : int  1744 2914 346525 301500 10974 15502 4997 53749 1975 198107 ...
##  $ user_time_zone            : logi  NA NA NA NA NA NA ...
##  $ user_urls                 : chr  NA NA NA "https://www.patreon.com/ShaunKronenfeld" ...
##  $ user_verified             : chr  "false" "false" "false" "false" ...
```

```
head(data1) # most informative
```

```
##   coordinates                     created_at         hashtags media
## 1        <NA> Sat May 16 04:13:37 +0000 2020  TheTourWillGoOn  <NA>
## 2        <NA> Sat May 16 04:13:35 +0000 2020             <NA>  <NA>
## 3        <NA> Sat May 16 04:13:36 +0000 2020             <NA>  <NA>
## 4        <NA> Sat May 16 04:13:35 +0000 2020 BREAKING COVID19  <NA>
## 5        <NA> Sat May 16 04:13:37 +0000 2020             <NA>  <NA>
## 6        <NA> Sat May 16 04:13:35 +0000 2020        HowWeFeel  <NA>
##                                                          urls favorite_count
## 1 https://twitter.com/kennychesney/status/1261463380809809920              0
## 2                                                        <NA>              0
## 3                                     https://yhoo.it/2WqISbc              0
## 4                                                        <NA>              0
## 5                                      https://bit.ly/3byxGh7              0
## 6                             https://get.howwefeel.org/share              0
##            id in_reply_to_screen_name in_reply_to_status_id in_reply_to_user_id
## 1 1.26151e+18                    <NA>                    NA                  NA
## 2 1.26151e+18                    <NA>                    NA                  NA
## 3 1.26151e+18                    <NA>                    NA                  NA
## 4 1.26151e+18                    <NA>                    NA                  NA
## 5 1.26151e+18                    <NA>                    NA                  NA
## 6 1.26151e+18                    <NA>                    NA                  NA
##   lang place possibly_sensitive     quote_id retweet_count   retweet_id
## 1   en  <NA>              false 1.261463e+18             0           NA
## 2   en  <NA>               <NA>           NA          7171 1.261250e+18
## 3   en  <NA>              false           NA            10 1.261509e+18
## 4   en  <NA>               <NA>           NA            76 1.261460e+18
## 5   en  <NA>              false           NA             9 1.261508e+18
## 6   en  <NA>              false           NA             0           NA
##   retweet_screen_name
## 1                <NA>
## 2       MSharifpourMD
## 3        YahooFinance
## 4          davejourno
## 5       IndianExpress
## 6                <NA>
##                                                                                  source
## 1    <a href="http://twitter.com/#!/download/ipad" rel="nofollow">Twitter for iPad</a>
## 2   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 3 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 4   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
## 5 <a href="http://twitter.com/download/android" rel="nofollow">Twitter for Android</a>
## 6   <a href="http://twitter.com/download/iphone" rel="nofollow">Twitter for iPhone</a>
##
## 1
## 2                   For those protesting social distancing (doubt they would read this) - a small gr
## 3
## 4 #BREAKING: Texas Supreme Court temporarily halts vote-by-mail expansion for #COVID19, siding tonig
## 5
## 6                                                                                      Iâ\200\231m u
##                                                         tweet_url
## 1   https://twitter.com/kevinhansford/status/1261510087874678784
## 2         https://twitter.com/pbm_ssb/status/1261510079486066688
## 3     https://twitter.com/chicago2503/status/1261510085064392705
## 4       https://twitter.com/bigred_13/status/1261510078684749824
## 5      https://twitter.com/khulibaat1/status/1261510087408881664
## 6 https://twitter.com/TommyCThompson4/status/1261510080459046913
##                  user_created_at      user_id user_default_profile_image
## 1 Thu May 14 15:39:20 +0000 2009 4.001776e+07                      false
## 2 Wed Oct 04 21:22:03 +0000 2017 9.156885e+17                      false
## 3 Fri May 01 06:18:42 +0000 2009 3.689734e+07                      false
## 4 Thu Feb 05 21:30:02 +0000 2009 2.019074e+07                      false
## 5 Wed Apr 02 11:28:08 +0000 2014 2.423660e+09                      false
## 6 Sun Apr 26 23:38:29 +0000 2020 1.254555e+18                      false
##
## 1 Huge sports fan. Alabama Football, Kentucky Basketball and Green Bay Packers are the teams I follow
## 2                                                            Smash Ult. player from NEOH. ðŸ\217³ï¸\
## 3
## 4
## 5
## 6
##   user_favourites_count user_followers_count user_friends_count
## 1                  8456                   52                446
## 2                 25418                   87                 95
## 3                   103                 1333               5002
## 4                156943                  586                499
## 5                 23317                  166                763
## 6                 12011                   83                206
##   user_listed_count user_location                           user_name
## 1                 5          <NA>             Kevin â\200œThe Senatorâ\200\235
## 2                 1     Formation             BLM | The little garage
## 3               158          <NA>                      paul@dodgerman
## 4                47          <NA> Rebecca Kronenfeld ðŸ\217³ï¸\217â\200\215âš§ï¸\217
## 5                 0          <NA>                         Askar Wasti
## 6                 2          <NA>                   Tommy C. Thompson
##   user_screen_name user_statuses_count user_time_zone
## 1    kevinhansford                1744             NA
## 2          pbm_ssb                2914             NA
## 3      chicago2503              346525             NA
## 4        bigred_13              301500             NA
## 5       khulibaat1               10974             NA
## 6  TommyCThompson4               15502             NA
##                                  user_urls user_verified
## 1                                     <NA>         false
## 2                                     <NA>         false
## 3                                     <NA>         false
## 4 https://www.patreon.com/ShaunKronenfeld         false
## 5                  http://www.alfaj-ar.com         false
## 6      HTTP://ToraAquilaDracos.Tripod.Com/         false
```

```
summary(data1) #not of much information as is mostly text in data set
```

```
##  coordinates         created_at          hashtags            media
##  Length:2274036     Length:2274036     Length:2274036     Length:2274036
##  Class :character    Class :character    Class :character    Class :character
##  Mode  :character    Mode  :character    Mode  :character    Mode  :character
```

```
##
##
##
##
##      urls           favorite_count           id           
##  Length:2274036     Min.   :     0.0   Min.   :1.262e+18  
##  Class :character   1st Qu.:     0.0   1st Qu.:1.262e+18  
##  Mode  :character   Median :     0.0   Median :1.262e+18  
##                     Mean   :     2.9   Mean   :1.262e+18  
##                     3rd Qu.:     0.0   3rd Qu.:1.262e+18  
##                     Max.   :413206.0   Max.   :1.262e+18  
##
##  in_reply_to_screen_name in_reply_to_status_id in_reply_to_user_id
##  Length:2274036          Min.   :2.631e+07     Min.   :1.200e+01  
##  Class :character        1st Qu.:1.262e+18     1st Qu.:4.364e+07  
##  Mode  :character        Median :1.262e+18     Median :4.327e+08  
##                          Mean   :1.261e+18     Mean   :2.653e+17  
##                          3rd Qu.:1.262e+18     3rd Qu.:7.293e+17  
##                          Max.   :1.262e+18     Max.   :1.262e+18  
##                          NA's   :2128821       NA's   :2108885    
##      lang              place           possibly_sensitive    quote_id        
##  Length:2274036     Length:2274036     Length:2274036     Min.   :4.135e+16  
##  Class :character   Class :character   Class :character   1st Qu.:1.261e+18  
##  Mode  :character   Mode  :character   Mode  :character   Median :1.261e+18  
##                                                           Mean   :1.260e+18  
##                                                           3rd Qu.:1.262e+18  
##                                                           Max.   :1.262e+18  
##                                                           NA's   :1774886    
##  retweet_count      retweet_id        retweet_screen_name    source         
##  Min.   :     0   Min.   :3.395e+17   Length:2274036      Length:2274036    
##  1st Qu.:     1   1st Qu.:1.261e+18   Class :character    Class :character  
##  Median :    32   Median :1.262e+18   Mode  :character    Mode  :character  
##  Mean   :  1442   Mean   :1.261e+18                                         
##  3rd Qu.:   537   3rd Qu.:1.262e+18                                         
##  Max.   :362842   Max.   :1.262e+18                                         
##                   NA's   :662759                                            
##      text            tweet_url         user_created_at       user_id         
##  Length:2274036     Length:2274036     Length:2274036     Min.   :1.700e+01  
##  Class :character   Class :character   Class :character   1st Qu.:2.449e+08  
##  Mode  :character   Mode  :character   Mode  :character   Median :2.228e+09  
##                                                           Mean   :3.680e+17  
##                                                           3rd Qu.:8.981e+17  
##                                                           Max.   :1.262e+18  
##
##  user_default_profile_image user_description   user_favourites_count
##  Length:2274036             Length:2274036     Min.   :      0      
##  Class :character           Class :character   1st Qu.:   3106      
##  Mode  :character           Mode  :character   Median :  18741      
##                                                Mean   :  59077      
##                                                3rd Qu.:  68201      
##                                                Max.   :2044647      
##
##  user_followers_count user_friends_count user_listed_count  user_location     
##  Min.   :       0     Min.   :      0    Min.   :     0.0   Length:2274036    
##  1st Qu.:     167     1st Qu.:    279    1st Qu.:     0.0   Class :character
##  Median :     645     Median :    821    Median :     3.0   Mode  :character
##  Mean   :   20077     Mean   :   2243    Mean   :   112.7
##  3rd Qu.:    2265     3rd Qu.:   2386    3rd Qu.:    23.0
##  Max.   :72132163     Max.   :1423338    Max.   :211459.0
##
##   user_name          user_screen_name   user_statuses_count user_time_zone
##  Length:2274036     Length:2274036     Min.   :      1     Mode:logical
##  Class :character   Class :character   1st Qu.:   6047     NA's:2274036
##  Mode  :character   Mode  :character   Median :  24933
##                                        Mean   :  77082
##                                        3rd Qu.:  79912
##                                        Max.   :7392635
##
##   user_urls          user_verified
##  Length:2274036     Length:2274036
##  Class :character   Class :character
##  Mode  :character   Mode  :character
##
##
##
##
```

```
#number of record that include a value for fields: user_location,coordinates,place: This fields could b
length(data1$user_location)-length(which(is.na(data1$user_location)))
```

```
## [1] 1600775
```

```
length(data1$coordinates)-length(which(is.na(data1$coordinates)))
```

```
## [1] 1332
```

```
length(data1$place)-length(which(is.na(data1$place)))
```

```
## [1] 23469
```

```
#Ti inspect the appropriateness for strata building

#print(data1$user_location) #best option as has least amount of NA, but needs to clean up list city, co
head(data1$user_location)
```

```
## [1] NA          "Formation" NA          NA          NA          NA
```

```
#print(data1$coordinates) #cleanest list with data points
head(data1$coordinates)
```

```
## [1] NA NA NA NA NA NA
```

```
#print(data1$place)
head(data1$place)
```

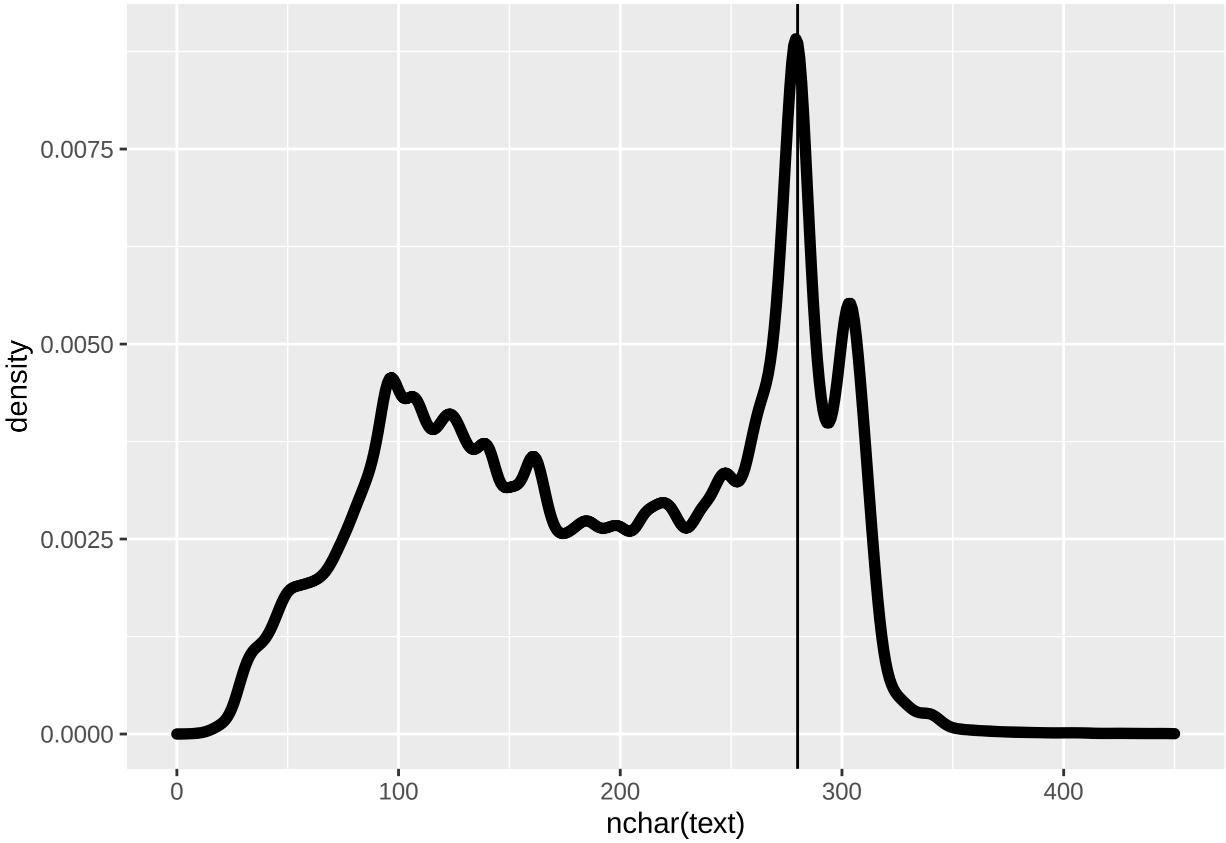
```
## [1] NA NA NA NA NA NA
```

```
#distribution of the number of characters in the data set attribute text / tweets content

ggplot(data = data1, aes(x = nchar(text))) + geom_density(size = 2) + geom_vline(xintercept = 280) + sca
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

```
## Warning: Removed 1763 rows containing non-finite values (stat_density).
```

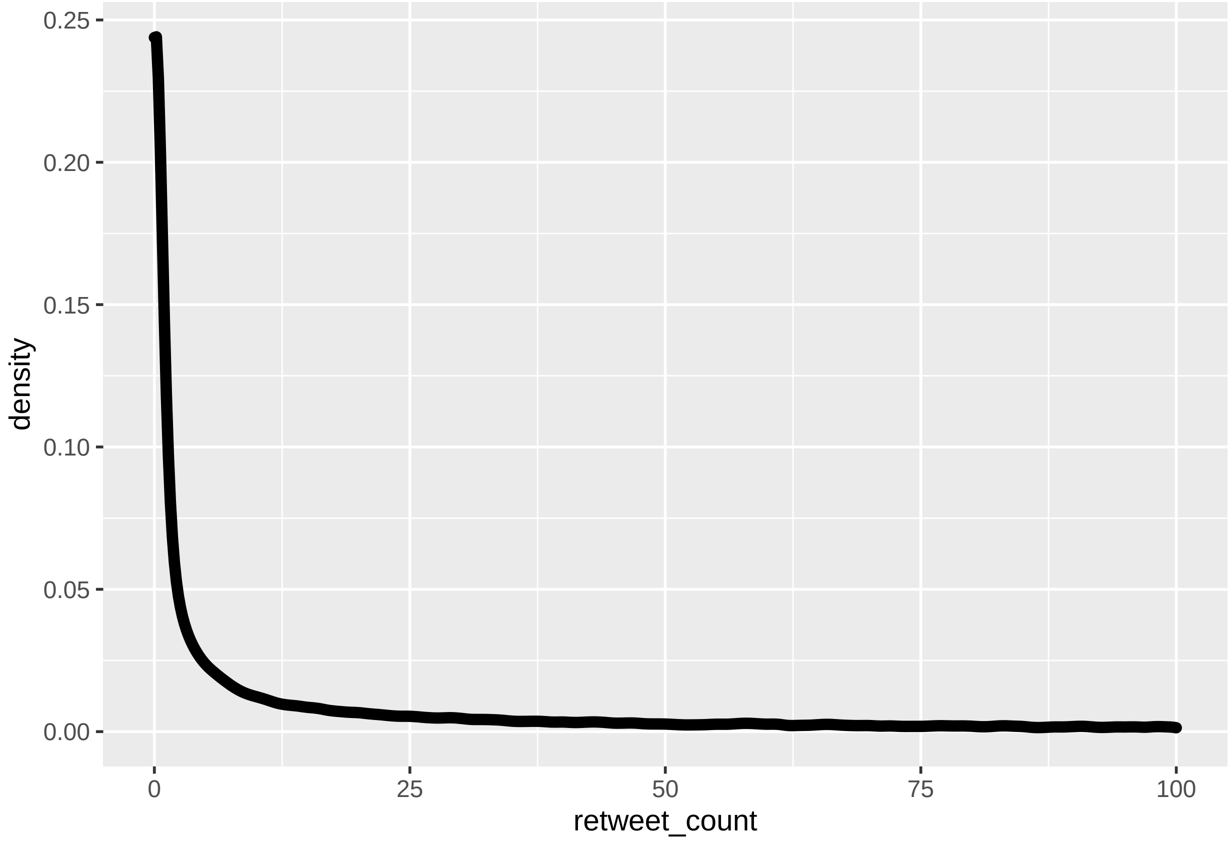


```
#This is a density graph : Computes and draws kernel density estimate, which is a smoothed version of t

#Conclusion: max number of characters per tweet is set at 280 by Twitter as can also been seen in the g

#Note: to remove scientific numbering , first create object p <- ggplot()
# p + scale_x_continuous(labels = function(x) format(x, scientific = FALSE))
# showing count of retweets in data set
ggplot(data = data1, aes(x = retweet_count)) + geom_density(size = 2) + xlim(0,100)
```

```
## Warning: Removed 914735 rows containing non-finite values (stat_density).
```



```
#Conslusion: only a few tweets are retweeted frequently.
```

```
#split attribute Coordinates into two columns
CoordinateDF <- data.frame(x = data1$coordinates)

SplitCoordinate <- CoordinateDF %>% separate(x, c("long","lat"), sep = "([,])")
```

```
#remove NAs
CoordinatesremoveNA <- na.omit(SplitCoordinate)

CoordinatesremoveNA$long <- as.numeric(CoordinatesremoveNA$long)
CoordinatesremoveNA$lat <- as.numeric(CoordinatesremoveNA$lat)
```

```
#building a world map of countries.
#Source: https://r-spatial.org/r/2018/10/25/ggplot2-sf.html#:~:text=This%20call%20nicely%20introduces%2

library(ggplot2)
theme_set(theme_bw())
library(sf)
```

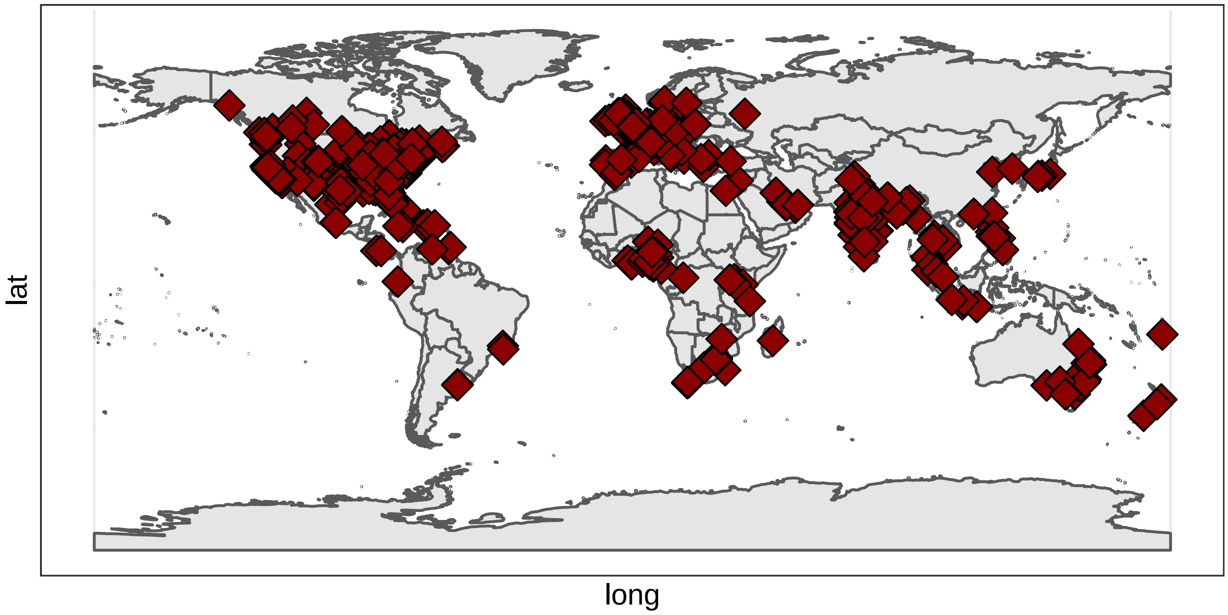
```
## Linking to GEOS 3.9.1, GDAL 3.2.1, PROJ 7.2.1
```

```
library("rnaturalearth")
library("rnaturalearthdata")

world <- ne_countries(scale = "medium", returnclass = "sf")
class(world)
```

```
## [1] "sf"         "data.frame"
```

```
#plotting data set to see geographical spread
ggplot(data = world) +
  geom_sf() +
  geom_point(data = CoordinatesremoveNA, aes(x = long, y = lat), size = 4,
             shape = 23, fill = "darkred")
```



```
# Zoom in by adding: + coord_sf(xlim = c(-88, -78), ylim = c(24.5, 33), expand = FALSE)

#save graph to PDF:
ggsave("map.pdf")
```

```
## Saving 6.5 x 4.5 in image
```

```
#show table with country names using photon


#install.packages('revgeo')

#library(devtools)
#install_github('mhudecheck/revgeo')

#library(revgeo)

#start <- Sys.time()
#This line do all the reverse geocoding using Photon as a provider
#results<-revgeo(longitude=CoordinatesremoveNA$long,
#                latitude=CoordinatesremoveNA$lat,
#                provider = 'photon', output="frame")

#end <- Sys.time()

#str(results)
```

```
#save object, results.
#saveRDS(results, file = "results.Rds")
```

```
#getwd()
#setwd("C:/Ryerson University - Capstone project/Module 2/EIEEE - Large dataset/Combined")
```

```
#load object results


results <- readRDS(file = "results.Rds")
str(results)
```

```
## ’data.frame’:    1332 obs. of  6 variables:
##  $ housenumber: chr  "718" "936" "House Number Not Found" "House Number Not Found" ...
##  $ street     : chr  "Orleans Avenue" "Shepherd Street Northwest" "Street Not Found" "Parkinson Way"
##  $ city       : chr  "New Orleans" "Washington" "Abuja" "Kelowna" ...
##  $ state      : chr  "Louisiana" "District of Columbia" "Federal Capital Territory" "British Columbia
##  $ zip        : chr  "70116" "20011" "900281" "V1Y6G2" ...
##  $ country    : chr  "United States" "United States" "Nigeria" "Canada" ...
```

```
#Create list frequency by city

install.packages("stats")
```

```
## Warning: package ’stats’ is in use and will not be installed
```

```
#aggregate(results$city, by=list(results$city), FUN=length)
res <- aggregate(results$city, by=list(results$city), FUN=length)
#head(res, 40)
#res[order(res$x, decreasing = TRUE),]
```

```
#Create a table Top-20
# save as dataframe, then plot frequency in ggplot
Locations <- data.frame(res[order(res$x, decreasing = TRUE),])
str(Locations)
```

```
## 'data.frame':    571 obs. of  2 variables:
##  $ Group.1: chr  "City Not Found" "New York" "Los Angeles" "London" ...
##  $ x      : int  175 90 41 34 20 18 15 14 12 11 ...
```

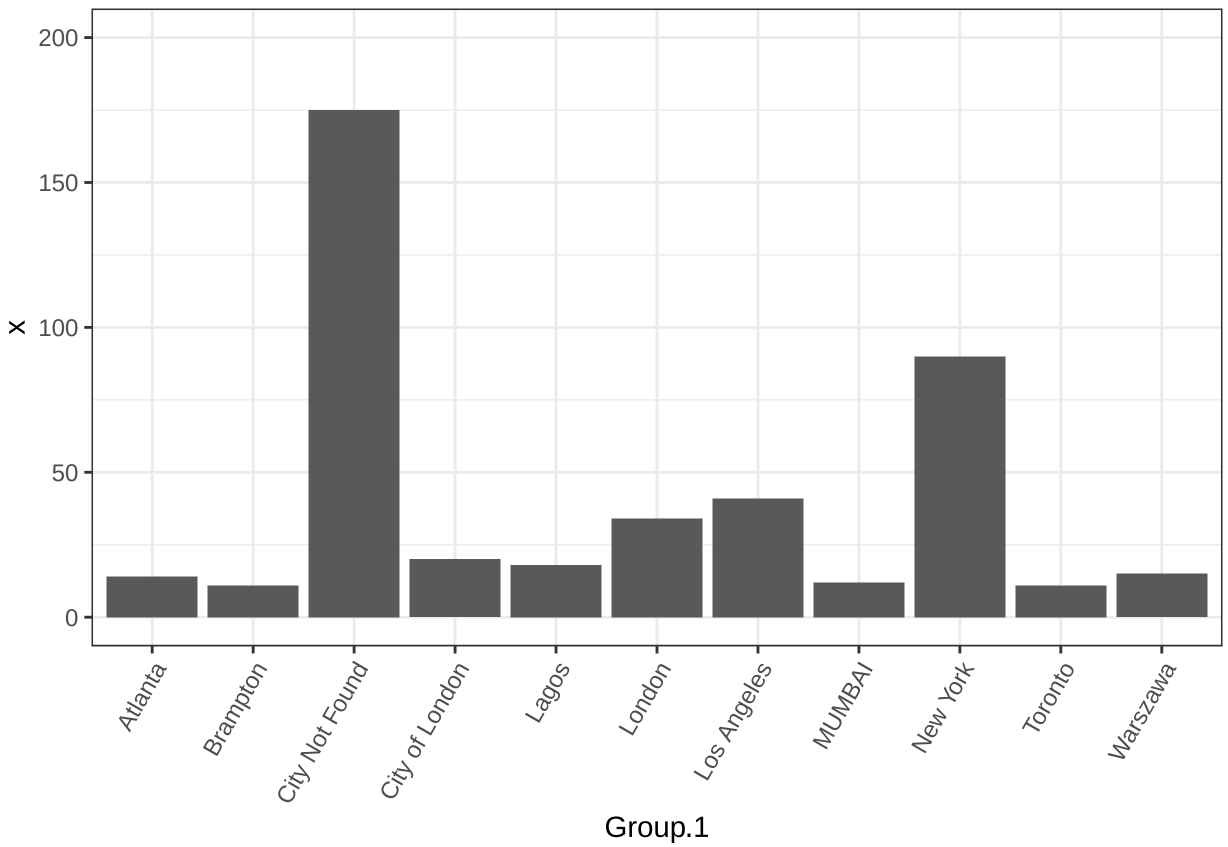
```
Locations$x = as.numeric(Locations$x)
length(Locations$x) #out of 1,332 coordinates (long,lat), only 571 returned with a city name including
```

```
## [1] 571
```

```
newdf <- subset(Locations, x > 10)
newdf
```

```
##            Group.1   x
## 135 City Not Found 175
## 371       New York  90
## 312    Los Angeles  41
## 309         London  34
## 136 City of London  20
## 286          Lagos  18
## 544       Warszawa  15
## 61         Atlanta  14
## 360         MUMBAI  12
## 96        Brampton  11
## 520        Toronto  11
```

```
ggplot(newdf,aes(x=Group.1, y=x)) + geom_bar(stat = 'identity') + scale_y_continuous(limits = c(0, 200)
```



```
#+ scale_x_discrete(name ='x')
```